Structural Databases of Biological Macromolecules

Helen M Berman, *Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA*

The Protein Data Bank began as an archive of the structural data available about known biological macromolecules. The advances made in all technologies have been mirrored in further development of the Protein Data Bank and in the structural, speciality and structural characteristic databases that have also evolved.

Advanced article

Article contents

doi: 10.1038/npg.els.0005252

Historical Background

In 1957, the first structure of a biological macromolecule (myoglobin) was determined (Kendrew *et al.*, 1958). This was followed by the determinations of several more key molecules, including hemoglobin (Perutz *et al.*, 1960), lysozyme (Blake *et al.*, 1965) and ribonuclease (Kartha *et al.*, 1967; Wyckoff *et al.*, 1967). In 1971, small-molecule and protein crystallographers from both sides of the Atlantic agreed to establish a data bank of the protein structures being determined. Its mission would be to collect, archive and disseminate data on the three-dimensional structures of biological macromolecules. Walter Hamilton of the Brookhaven National Laboratory and Olga Kennard of the Cambridge Structural Database (CSD) collaborated to manage the Protein Data Bank (PDB) resource (1971). Hamilton's interest was borne from his work on the high-resolution determination of amino acid crystal structures and from his visionary idea of setting up distributed computing resources whereby every crystallographer would have a graphics workstation on his/her desk with full network access to powerful high-speed computers. Kennard had founded the CSD in 1965 to create a database of organic and metal-organic compounds studied by X-ray and neutron diffraction, and was well experienced in managing structural data. (***See*** Crystallization of Nucleic Acids; Protein Structure.)

The PDB contained less than a dozen structures at its inception, with a few more structures added each year. The structures themselves were relatively small. The PDB file format was simple, and it was relatively easy to extract the structures from magnetic tape to find out what you wanted to know about any particular molecule.

In the 1980s, the improvements in the technology required to do crystal structures began to evolve rapidly. Now, two decades later, modern molecular biology techniques have made it much more straightforward to obtain large quantities of proteins. Crystallization methods have emerged that allow investigators to screen many different conditions using exceedingly small amounts of material. Data collection methods have improved at all levels. The lifetimes of crystals are routinely extended by flash freezing. The radiation sources are much more intense, especially with the emergence of powerful synchrotron beam lines. Detectors are much more sensitive and allow the very rapid collection of arrays of reflections. Methods for phase determination and refinement have improved. Indeed, crystallography is now part of the armament of techniques that is readily accessible to biologists.

As crystallographic methods continue to improve, another method of structure determination has come of age: nuclear magnetic resonance (NMR). This method, which allows the determination of structures in solution, is currently responsible for approximately 15% of the structures released in the PDB.

The improvements in technology have also made it possible to determine the structures of very complex molecules. Several structures of ribosomal subunits (Moore, 2001), as well as the entire 70S ribosome structure (Yusupov *et al.*, 2001), have been deposited in the PDB. During this same period, the structural genomics initiative (2000) has begun with the goal of determining thousands of structures in a high-throughput mode. Thus, the PDB holdings will continue to grow (**Figure 1**).

The level of activity in structural biology has made it essential that the PDB use the most modern technologies to collect, archive and disseminate data. The PDB is an *archival database*, which contains coordinates of biological macromolecules determined using public funds as well as many from the private sector. It also contains information about the methods and materials used to determine those structures. Other databases have emerged (**Table 1**) that extract some of the information contained in the PDB and

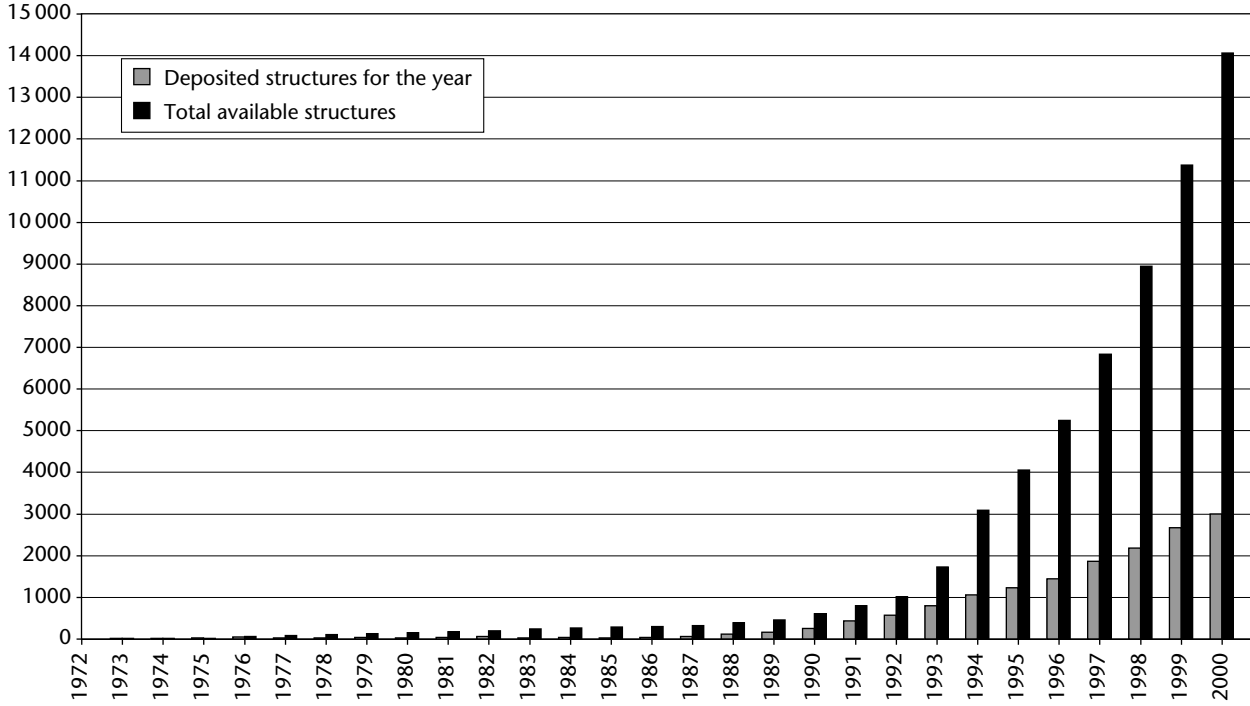

Figure 1 Growth of the contents of the Protein Data Bank. The number of structures deposited each year is shown in gray, the total number of structures available in black. This chart is regularly updated at http://www.rcsb.org/pdb/holdings.html.

organize that information in different ways so as to enable different types of query. These are *value-added databases*, which serve the needs of particular users. In this article we describe the PDB and some of these other structural databases.

The Protein Data Bank

After 27 years at Brookhaven National Laboratory, the PDB is now managed by the Research Collaboratory of Structural Bioinformatics (RCSB) (Berman *et al*., 2000). The RCSB is a consortium consisting of three member groups: Rutgers, the State University of New Jersey; the San Diego Supercomputer Center of the University of California, San Diego; and the National Institute of Standards and Technology. The PDB collects information about biological macromolecular structures and the methods used to determine those structures. Coordinates, primary experimental data, statistics about the structure determination and refinement, information about the source, sequence and chemistry of the molecule and the solution and/or crystallization conditions are collected and assembled using a software tool called the AutoDep Input Tool (ADIT) (see Web Links). Annotation, checking and validation of the data are carried out with a variety of programs whose output is reviewed by skilled annotation staff working in close collaboration with the depositors of the data.

Once the data are fully checked and approved for release, they are loaded into a series of databases. Two different search engines can query the databases: *SearchLite* and *SearchFields*. A rich set of reporting options make it possible to access information about a single molecule, compare it with other molecules, and access other databases that contain information about that molecule. Particular groups of macromolecules can be selected according to their features so that a variety of reports can be created. The PDB maintains mirrors around the world, which provide the same capabilities as the main RCSB site.

As of this writing there are more than 15 800 molecules in the PDB. The distribution of these data is shown in **Table 2**.

Structural Databases

While the PDB focuses on individual structures, some databases organize their data according to tertiary structural characteristics. SCOP (a Structural Classification of Proteins) classifies each structure in the PDB according to *family*, *superfamily*, *common fold* and *class*. *Families* are classified according to their sequence similarities. Families with similar structure and function belong to the same *superfamily*. Families and

Table 1 Selected database resources for macromolecular structures

Archival database of biological macromolecules	
Protein Data Bank (Berman *et al*., 2000; Bernstein *et al*., 1977)	http://www.pdb.org/
Structural databases	
3D ALI (a database of aligned protein structures and related sequences) (Pascarella and Argos, 1992)	http://www.embl-heidelberg.de/argos/ali/ali_info.html
CAMPASS (Sowdhamini *et al*., 1998)	http://www-cryst.bioc.cam.ac.uk/~campass/
CATH (Orengo *et al*., 1997)	http://www.biochem.ucl.ac.uk/bsm/cath/
CSD (Allen *et al*., 1979)	http://www.ccdc.cam.ac.uk/
FSSP (Holm and Sander, 1998)	http://www2.ebi.ac.uk/dali/fssp/
HSSP (Dodge *et al*., 1998)	http://www.sander.embl-heidelberg.de/hssp/
ISSD (Adzhubei *et al*., 1998)	http://www.protein.bio.msu.su/issd/
Library of Protein Family Cores (LPFC) (Schmidt *et al*., 1997)	http://WWW-SMI.Stanford.EDU/projects/helix/LPFC/
Molecular Modeling Database (Holm and Sander, 1994)	http://www.ncbi.nlm.nih.gov/Structure/
SCOP (Murzin *et al*., 1995)	http://scop.mrc-lmb.cam.ac.uk/scop/
Speciality databases	
ENZYME database (Bairoch, 2000)	http://www.expasy.ch/enzyme/
Enzyme Structures Database	http://www.biochem.ucl.ac.uk/bsm/enzymes/
HIV Protease Database (Vondrasek *et al*., 1997)	http://srdata.nist.gov/hivdb/
International Immunogenetics Database (IMGT) (Lefranc *et al*., 1998)	http://imgt.cines.fr:8104/
Nucleic Acid Database (Berman *et al*., 1992)	http://ndbserver.rutgers.edu/
Prolysis (protease and protease inhibitor web server)	http://delphi.phys.univ-tours.fr/Prolysis/
Protein Kinase Resource (Smith *et al*., 1997)	http://pkr.sdsc.edu/html/index.shtml
Structural characteristic databases	
Biological Macromolecule Crystallization Database (BMCD) (Gilliland, 1997)	http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html
Dictionary of Interfaces in Proteins (DIP)	http://www.drug-redesign.de/
ISOSTAR (Bruno *et al*., 1997)	http://www.ccdc.cam.ac.uk/prods/isostar/
Molecular Movements Database (Gerstein and Krebs, 1998)	http://bioinfo.mbb.yale.edu/MolMovDB/
OLDERADO (Kelley and Sutcliffe, 1997)	http://neon.chem.le.ac.uk/olderado/
PDBSum (Laskowski *et al*., 1997)	http://www.biochem.ucl.ac.uk/bsm/pdbsum/
PROCAT (Wallace *et al*., 1996)	http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html
PROMISE (Degtyarenko *et al*., 1998)	http://metallo.scripps.edu/PROMISE/
Protein Quaternary Structures (PQS)	http://pqs.ebi.ac.uk/
ReLiBase (Receptor/ligand complexes database) (Hendlich *et al*., 2003)	http://relibase.ccdc.cam.ac.uk/
TESS	

Table 2 Protein Data Bank holdings (as of 14 August 2001)

	Proteins, peptides and viruses	Protein–nucleic acid complexes	Nucleic acids	Carbohydrates	Total
X-ray diffraction and other	11 893	569	579	14	13 055
NMR	1964	73	390	4	2431
Theoretical modeling	293	20	23	0	336
Total	14 150	662	992	18	15 822

superfamilies with the same arrangement of secondary structures, which are connected with one another in the same way, have the same *common fold*. *Class* refers to the types of secondary structures (all alpha, all beta, alpha–beta, etc.). SCOP was one of the earliest databases that attempted to integrate sequence, structure and function information; it continues to be a major resource in structural biology.

CATH provides another classification scheme based on class (C), architecture (A), topology (T) and homologous superfamilies (H). *Class* defines the secondary structure content as in SCOP. *Architecture* defines the description of the arrangement of these secondary structures without consideration of the connectivities. *Topology* is equivalent to fold in SCOP. Finally, *homologous superfamilies* contain all folds with a similar function. CATH has a systematic classification system for all structures analogous to the EC classification for enzyme function. The type of research possible with this database is exemplified by an analysis of all enzymes in which it was shown that the topology of enzymes is more related to the ligands bound than the enzyme EC class (Martin *et al*., 1998).

Speciality Databases

Another type of database that has proved invaluable in research has been the speciality database. These databases are curated by experts in the field and provide information beyond the structures themselves. These may include derived structural data, sequence information and other biochemical information. An example is the Protein Kinase Resource, which provides not only structural but also functional and pharmacological data about these key drug targets. The same is true of the HIV Protease Database, which has captured all the information about HIV protease structures to be included in one place with the goal of aiding drug development. The Nucleic Acid Database (NDB) has provided a searchable resource about nucleic acids. The Enzyme Structures Database organizes all the enzyme structures in the PDB according to the EC codes contained in the ENZYME Data Bank and provides information about them. (***See*** DNA Structure.)

Databases of Structural Characteristics

Databases of structural features contained within macromolecules have also emerged. The Molecular Movements Database provides information about the possible motions of macromolecules by analyzing the various structures of particular molecules. OLDERADO (On Line Database of Ensemble Representatives And Domains) provides a database of structures for which there are several representatives, such as an ensemble of NMR structures. The TESS (Template Search and Superimposition) algorithm has allowed for the creation of a database of active site templates called PROCAT. This type of database will become invaluable in the quest for relating structure to function. PROMISE is a database that provides information about the prosthetic centers and metal ions in the active sites. ISOSTAR provides an integrated view of the nonbonded interactions geometry around ligands in proteins. PDBsum gives a variety of carefully curated information about all the structures in the PDB. The Dictionary of Interfaces in Proteins (DIP) is a data bank of complementary molecular surface patches and is meant to enable molecular recognition research.

Challenges

The PDB is now much more than a repository of coordinate data. To make this resource even more useful, all the files need to be in a uniform format so that the many new databases of derived information can be easily constructed without having to first clean the files. A project at the PDB is underway to re-examine the archive to achieve this uniformity (Bhat *et al*., 2001). The PDB will also integrate the validation criteria that have been developed by a variety of researchers (Wilson *et al*., 1998).

The various methods that have been developed for classification (Gerstein and Levitt, 1998; Orengo and Taylor, 1996) and structure comparison (Alexandrov and Fischer, 1996; Gibrat *et al*., 1996; Holm and Sander, 1996; Shindyalov and Bourne, 1998) will continue to improve and their results incorporated into the databases, as will methods to understand macromolecular interactions with one another (Jones and Thornton, 1997), with nucleic acids (Jones *et al*., 1999) and with small molecule ligands (Wallace *et al*., 1995).

The goal of being able to relate structure to function will be facilitated by different types of database efforts. Databases that assemble information about particular protein families will be one avenue that will provide this information. In these databases the coverage is very narrow and deep, so that a truly full understanding of a single class of proteins with known function is possible. The lessons learned from these types of resource will perhaps allow us to develop some general principles about the relationships of structure and function.

The structural genomics project is an outgrowth of the various genome projects. Its goal is to determine macromolecular structures on a genomic scale – the discovery, analysis and dissemination of three-dimensional structures of biological macromolecules representing the entire range of structural diversity found in nature (see Web Links). The sequences being targeted by many of these efforts are being stored in a database (see Web Links). Once the anticipated large volume of three-dimensional data is collected and assembled, it will be critical to coordinate and to relate

the structural and sequence data in order to create a full picture of protein fold space.

While these efforts are ongoing, databases of information about chemical and biological properties of macromolecules and their complexes will provide yet another avenue to understanding function.

With the large number of databases that have been created, it is important to develop methods to query across all of these databases in a seamless way. To help in this effort, the RCSB has developed a standard application interface for macromolecular data based on the Common Object Request Broker Architecture (Corba). The proposal was adopted by the Object Management Group (OMB) in February 2001 (see Web Links). This specification opens the door to more seamless and specific access to PDB data. More specifically, it provides a standard application programming interface (API) that will allow direct access by remote programs to the binary data structures of the PDB. This and other similar initiatives will help to ensure that the world of biology *in silico* will be readily accessible.

Acknowledgements

Parts of this work have appeared in 'The past and future of structural databases' by Helen M. Berman (1999) *Current Opinion in Biotechnology* **10**: 76–80. The US National Science Foundation, the Department of Energy and the National Institutes of Health supported this work.

See also

Protein Databases

References

(1971) Protein Data Bank. *Nature New Biology* **233**: 223.

Adzhubei IA, Adzhubei AA and Neidle S (1998) An integrated sequence-structure database incorporating matching mRNA sequence and protein three-dimensional structure data. *Nucleic Acids Research* **26**: 327–331.

Alexandrov NN and Fischer D (1996) Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. *Proteins* **25**: 354–365.

Allen FH, Bellard S, Brice MD, *et al*. (1979) The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallographica B* **35**: 2331–2339.

Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Research* **28**: 304–305.

Berman HM, Olson WK, Beveridge DL, *et al*. (1992) The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal* **63**: 751–759.

Berman HM, Westbrook J, Feng Z, *et al*. (2000) The Protein Data Bank. *Nucleic Acids Research* **28**: 235–242.

Bernstein FC, Koetzle TF, Williams GJB, *et al*. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology* **112**: 535–542.

Bhat TN, Bourne P, Feng Z, *et al*. (2001) The PDB data uniformity project. *Nucleic Acids Research* **29**: 214–218.

Blake CC, Koenig DF, Mair GA, *et al*. (1965) Structure of hen egg-white lysozyme. A three dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**: 757–761.

Bruno IJ, Cole JC, Lommerse JP, *et al*. (1997) ISOSTAR: a library of information about nonbonded interactions. *Journal of Computer-Aided Molecular design* **11**: 525–537.

Degtyarenko KN, North ACT, Perkins DN and Findlay JBC (1998) PROMISE: a database of information on prosthetic centres and metal ions in protein active sites. *Nucleic Acids Research* **26**: 376–381.

Dodge C, Schneider R and Sander C (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Research* **26**: 313–315.

Gerstein M and Krebs W (1998) A database of macromolecular motions. *Nucleic Acids Research* **26**: 4280–4290.

Gerstein M and Levitt M (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Science* **7**: 445–456.

Gibrat J-F, Madej T and Bryant SH (1996) Surprising similarities in structure comparison. *Current Opinions in Structural Biology* **6**: 377–385.

Gilliland GL (1997) Biological Macromolecule Crystallization Database. *Methods in Enzymology* **277**: 546–556.

Hendlich M, Bergner A, Günther J and Klebe G (2003) ReLiBase: Design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of Molecular Biology* **326**: 607–620.

Holm L and Sander C (1994) Searching protein structure databases has come of age. *Proteins* **19**: 165–173.

Holm L and Sander C (1996) Mapping the protein universe. *Science* **273**: 595–603.

Holm L and Sander C (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Research* **26**: 316–319.

Jones S, van Heyningen P, Berman HM and Thornton JM (1999) Protein–DNA interactions: a structural analysis. *Journal of Molecular Biology* **287**: 877–896.

Jones S and Thornton JM (1997) Analysis of protein–protein interaction sites using surface patches. *Journal of Molecular Biology* **272**: 121–132.

Kartha G, Bello J and Harker D (1967) Tertiary structure of ribonuclease. *Nature* **213**: 862–865.

Kelley LA and Sutcliffe MJ (1997) OLDERADO: On-line database of ensemble representatives and domains. *Protein Science* **6**: 2628–2630.

Kendrew JC, Bodo G, Dintzis HM, Parrish RG and Wyckoff H (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**: 662–666.

Laskowski RA, Hutchinson EG, Michie AD, *et al*. (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences* **22**: 488–490.

Lefranc MP, Giudicelli V, Busin C, *et al*. (1998) IMGT, the International ImMunoGeneTics database. *Nucleic Acids Research*, **26**: 297–303.

Martin ACR, Orengo CA, Hutchinson EG, *et al*. (1998) Protein folds and functions. *Structure* **6**: 875–884.

Moore P (2001) The ribosome at atomic resolution. *Biochemistry* **40**: 3243–3250.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation

of sequences and structures. *Journal of Molecular Biology* **247**: 536–540.
Orengo CA, Michie AD, Jones S, *et al*. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
Orengo CA and Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. *Methods in Enzymology* **266**: 617–635.
Pascarella S and Argos P (1992) Analysis of insertions/deletions in protein structures. *Journal of Molecular Biology* **224**: 461–471.
Perutz MF, Rossmann MG, Cullis AF, Muirhead G and Will G (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution. *Nature* **185**: 416–422.
Schmidt R, Gerstein M and Altman R (1997) LPFC: An Internet library of protein family core structures. *Protein Science* **6**: 246–248.
Shindyalov IN and Bourne PE (1998) Protein structure alignment by incremental combinatory extension of the optimum path. *Protein Engineering* **11**: 739–747.
Smith C, Gribskov M, Shindyalov IN, *et al*. (1997) The Protein Kinase Resource. *Trends in Biochemical Sciences* **22**: 444–446.
Sowdhamini R, Burke DF, Huang J-f, *et al*. (1998) CAMPASS: a database of structurally aligned protein. *Structure* **6**: 1087–1094.
Vondrasek J, Buskirk C and Wlodawer A (1997) Database of three-dimensional structures of HIV proteinases. *Nature Structural Biology* **4**: 8.
Wallace A, Laskowski R and Thornton J (1996) Derivation of 3D coordinate templates for searching structural databases: application to the Ser-His-Asp catalytic triads of the serine proteinases and lipases. *Protein Science*, **5**: 1001–1013.
Wallace AC, Laskowski RA and Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein ligand interactions. *Protein Engineering* **8**: 127–134.
Wilson KS, Butterworth S, Dauter Z, *et al*. (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *Journal of Molecular Biology* **276**: 417–436.
Wyckoff HW, Hardman KD, Allewell NM, *et al*. (1967) The structure of ribonuclease-S at 6 Å resolution. *Journal of Biological Chemistry* **242**: 3749–3753.
Yusupov MM, Yusupova GZ, Baucom A, *et al*. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science* **282**: 883–896.

Further Reading

(2000) Structural genomics supplement. *Nature Structural Biology* **7**: 927–994 [entire issue].
(2001) Database Issue. *Nucleic Acids Research* **29**: 1–349.
(2001) *The Structures of Life*. NIH publication number 01–2778. http://www.nigms.nih.gov/news/science_ed/structlife.pdf.
Benton D (1996) Bioinformatics – principles and potential of a new multidisciplinary tool. *Trends in Biotechnology* **14**: 261–272.
Berman HM, Bhat TN, Bourne PE, *et al*. (2000) The Protein Data Bank and the challenge of structural genomics. *Nature Structural Biology* **7**: 957–959.
Berman HM, Gelbin A and Westbrook J (1996) Nucleic acid crystallography: a view from the Nucleic Acid Database. *Progress in Biophysics and Molecular Biology* **66**: 255–288.
Gaasterland T (1998) Structural genomics: bioinformatics in the driver's seat. *Nature Biotechnology* **16**: 625–627.
Holm L and Sander C (1994) Searching protein structure databases has come of age. *Proteins* **19**: 165–173.
Rost B (1998) Marrying structure and genomics. *Structure* **6**: 259–263.
Swindells MB, Orengo CA, Jones DT, Hutchinson EG and Thornton JM (1998) Contemporary approaches to protein structure classification. *BioEssays* **20**: 884–891.
Westbrook J and Bourne PE (2000) STAR/mmCIF: an extensive ontology for macromolecular structure and beyond. *Bioinformatics* **16**: 159–168.
Wilson KS, Butterworth S, Dauter Z, *et al*. (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *Journal of Molecular Biology* **276**: 417–436.
Zou J-Y and Mowbray SL (1994) An evaluation of the use of databases in protein structure refinement. *Acta Crystallographica D* **50**: 237–249.

Web Links

PDB Deposition Information. Links to the AutoDep Input Tool (ADIT), AutoDep, and other deposition resources
http://www.pdb.org/
Second International Structural Genomics Meeting. NIGMS statement on coordinate deposition, highlights, agreed principles and procedures, roster, agenda, and Task Force Reports
http://www.nigms.nih.gov/news/meetings/airlie.html
TargetDB. Target Registration Database that contains sequences from the worldwide structural genomics centers, and the PDB
http://targetdb.pdb.org/
OMG/LSR Corba Standard for Macromolecular Structure Data (OMG specification formal/02-05-01). First formal version of the Macromolecular Structure specification
http://www.omg.org/technology/documents/formal/macro_molecular.htm
The OpenMMS Toolkit. Corba, Relation Database and XML Software for Macromolecular Structure
http://openmms.sdsc.edu/